

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

JCS600 U.S. PRO
05/871272
05/31/01

別紙添付の書類に記載されている事項は下記の出願書類に記載
る事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed
this Office.

願 年 月 日
Date of Application:

2000年 6月29日

願 番 号
Application Number:

特願2000-197421

願 人
Applicant(s):

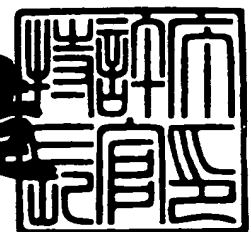
株式会社エス・エス・アール
学校法人高知工科大学

CERTIFIED COPY OF
PRIORITY DOCUMENT

2000年12月 1日

特 許 庁 長 官
Commissioner,
Patent Office

及 川 耕 造



【書類名】 特許願

【整理番号】 PB06007

【提出日】 平成12年 6月29日

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 17/30

【発明の名称】 テキストマイニングにおける文書の特徴量抽出方法及び
その装置

【請求項の数】 11

【発明者】

 【住所又は居所】 高知県南国市蛸が丘 1 - 1 - 1 株式会社エス・エス・
 アール内

 【氏名】 吉岡 倍達

【発明者】

 【住所又は居所】 高知県香美郡土佐山田町宮ノ口 1 8 5 高知工科大学情
 報システム工学科内

 【氏名】 ラック ターウォンマット

【特許出願人】

 【識別番号】 300044838

 【氏名又は名称】 株式会社エス・エス・アール

【特許出願人】

 【識別番号】 597154966

 【氏名又は名称】 学校法人高知工科大学

【代理人】

 【識別番号】 100077481

 【弁理士】

 【氏名又は名称】 谷 義一

【選任した代理人】

 【識別番号】 100088915

 【弁理士】

【氏名又は名称】 阿部 和夫

【選任した代理人】

【識別番号】 100106998

【弁理士】

【氏名又は名称】 橋本 傳一

【手数料の表示】

【予納台帳番号】 013424

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 0008215

【ブルーフの要否】 要

【書類名】 明細書

【発明の名称】 テキストマイニングにおける文書の特徴量抽出方法及びその装置

【特許請求の範囲】

【請求項 1】 文書の内容を代表する索引語に対応するベクトルからなる単語－文書行列を用いて前記文書の特徴量を抽出するテキストマイニングにおける文書の特徴量抽出方法であって、前記単語－文書行列の各要素には前記索引語に対する寄与分が作用し、コストを最小化する最急降下法に基いて互いに関連した文書および単語が近接する前記特徴量の空間を張る基底ベクトルを計算する基底ベクトル計算ステップと、前記単語－文書行列及び前記基底ベクトルを用いて前記特徴量を正規化するためのパラメータを計算し、該パラメータに基き前記特徴量を抽出する特徴量抽出ステップと、前記単語－文書行列を更新して前記基底ベクトルを適用しない前記単語－文書行列と適用した前記単語－文書行列との差分にする単語－文書行列更新ステップとを備えたことを特徴とするテキストマイニングにおける文書の特徴量抽出方法。

【請求項 2】 前記コストは、前記基底ベクトルを適用しない前記単語－文書行列と適用した前記単語－文書行列との差分の二次コストとして定義されることを特徴とする請求項 1 に記載のテキストマイニングにおける文書の特徴量抽出方法。

【請求項 3】 前記基底ベクトル計算ステップは、前記基底ベクトルの値を初期化する初期化ステップと、前記基底ベクトルの値を更新する基底ベクトル更新ステップと、前記基底ベクトルの値の変化度合いを求める変化度合い計算ステップと、前記基底ベクトルの値の変化度合いを用いて繰り返し処理を終了するかどうかを判別する判別ステップと、前記繰り返し処理の回数を数える計数ステップとを備えたことを特徴とする請求項 1 または 2 に記載のテキストマイニングにおける文書の特徴量抽出方法。

【請求項 4】 前記基底ベクトル更新ステップは、前記基底ベクトルの現在値と、前記単語－文書行列と、前記基底ベクトルの更新度合いを制御する更新率とを用いて前記基底ベクトルを更新することを特徴とする請求項 3 に記載のテキ

ストマイニングにおける文書の特徴量抽出方法。

【請求項 5】 前記特徴量の抽出に必要とされる全ての前記基底ベクトル及び前記正規化パラメータを既に取得している場合は、前記基底ベクトル計算ステップ及び前記特徴量抽出ステップにおける前記正規化パラメータの計算を省略し、前記特徴量抽出ステップは、既に取得している前記基底ベクトル及び前記正規化パラメータを用いて前記特徴量を抽出することを特徴とする請求項 1～4 のいずれか 1 項に記載のテキストマイニングにおける文書の特徴量抽出方法。

【請求項 6】 文書の内容を代表する索引語に対応するベクトルからなる単語－文書行列を用いて前記文書の特徴量を抽出するテキストマイニングにおける文書の特徴量抽出装置であって、前記単語－文書行列の各要素には前記索引語に対する寄与分が作用し、コストを最小化する最急降下法に基いて互いに関連した文書および単語が近接する前記特徴量の空間を張る基底ベクトルを計算する基底ベクトル計算手段と、前記単語－文書行列及び前記基底ベクトルを用いて前記特徴量を正規化するためのパラメータを計算し、該パラメータに基き前記特徴量を抽出する特徴量抽出手段と、前記単語－文書行列を更新して前記基底ベクトルを適用しない前記単語－文書行列と適用した前記単語－文書行列との差分にする単語－文書行列更新手段とを備えたことを特徴とするテキストマイニングにおける文書の特徴量抽出装置。

【請求項 7】 前記コストは、前記基底ベクトルを適用しない前記単語－文書行列と適用した前記単語－文書行列との差分の二次コストとして定義されることを特徴とする請求項 6 に記載のテキストマイニングにおける文書の特徴量抽出装置。

【請求項 8】 前記基底ベクトル計算手段は、前記基底ベクトルの値を初期化する初期化手段と、前記基底ベクトルの値を更新する基底ベクトル更新手段と、前記基底ベクトルの値の変化度合いを求める変化度合い計算手段と、前記基底ベクトルの値の変化度合いを用いて繰り返し処理を終了するかどうかを判別する判別手段と、前記繰り返し処理の回数を数える計数手段とを備えたことを特徴とする請求項 6 または 7 に記載のテキストマイニングにおける文書の特徴量抽出装置。

【請求項 9】 前記基底ベクトル更新手段は、前記基底ベクトルの現在値と、前記単語一文書行列と、前記基底ベクトルの更新度合いを制御する更新率とを用いて前記基底ベクトルを更新することを特徴とする請求項 8 に記載のテキストマイニングにおける文書の特徴量抽出装置。

【請求項 10】 前記特徴量の抽出に必要とされる全ての前記基底ベクトル及び前記正規化パラメータを既に取得している場合は、前記基底ベクトル計算手段及び前記特徴量抽出手段における前記正規化パラメータの計算を省略し、前記特徴量抽出手段は、既に取得している前記基底ベクトル及び前記正規化パラメータを用いて前記特徴量を抽出することを特徴とする請求項 6～9 のいずれか 1 項に記載のテキストマイニングにおける文書の特徴量抽出装置。

【請求項 11】 文書の内容を代表する索引語に対応するベクトルからなる単語一文書行列を用いて前記文書の特徴量を抽出するテキストマイニングにおける文書の特徴量抽出装置において実行される特徴量抽出プログラム・プロダクトであって、前記単語一文書行列の各要素には前記索引語に対する寄与分が作用し、コストを最小化する最急降下法に基いて互いに関連した文書および単語が近接する前記特徴量の空間を張る基底ベクトルを計算する基底ベクトル計算ステップと、前記単語一文書行列及び前記基底ベクトルを用いて前記特徴量を正規化するためのパラメータを計算し、該パラメータに基き前記特徴量を抽出する特徴量抽出ステップと、前記単語一文書行列を更新して前記基底ベクトルを適用しない前記単語一文書行列と適用した前記単語一文書行列との差分にする単語一文書行列更新ステップとを備えたことを特徴とする特徴量抽出プログラム・プロダクト。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、テキストマイニングにおける文書の特徴量抽出方法及びその装置に関し、より詳細には、特徴量を用いて文書および／またはウェブ検索、関連語検索、文書分類等の応用としてテキストマイニングを行う場合に、特徴量の空間において互いに関連した文書や単語が近接する特徴量を抽出するテキストマイニングにおける文書の特徴量抽出方法及びその装置に関する。

【0002】

【従来の技術】

文章データを種々の観点から分析し、所望の知識や情報を取り出す技術であるテキストマイニングにおいて、文書の有効な特徴量抽出は文書および／またはウェブ検索、関連語検索、文書分類などを効率よく行うための重要な課題である。一般的な文書の特徴量抽出方法としては、「Automatic Text Processing」(Addison-Wesley社、1989年出版)の第313項で述べられているベクトル空間法(vector-space model)がよく用いられている。

【0003】

ベクトル空間法では、文書の中で索引として選ばれた単語、即ち文書の内容を代表する索引語が t 個ある場合、それぞれの索引語 T_i にベクトル V_i を対応させ、 t 次元のベクトル空間を定義する。このように定義されたベクトル空間を構成する全てのベクトルは、 t 個の索引語に対応する t 個のベクトルの線形結合として表現できる。このベクトル空間において、文書 D_r を以下のように表現する。

【0004】

【数1】

$$D = \sum_{i=1}^t x_{ir} V_i \quad \text{式(1)}$$

【0005】

式(1)において、 V_i に作用する x_{ir} は文書 D_r における索引語 T_i に対する寄与分であり、文書 D_r の特徴量を表す。特徴量とは、索引語の各文書における出現頻度を表す量である。 $t \times 1$ (t 行1列)のベクトル $[x_{r1}, x_{r2}, \dots, x_{rt}]'$ は文書 D_r の特徴量ベクトルとなる。最も単純な場合としては、文書 D_r において索引語 T_i が出現する場合には1とし、出現しない場合には0とする方法がとられる。より複雑な場合は、上記の文献の第279項から第280項までで述べられているように、文書 D_r における索引語 T_i の出現頻度(term frequency) $t f_{ri}$ や、文書データベースに登録された全文書における索引語 T_i を含む文書頻度 $d f_i$ が x_{ir} の計算に利用される。

【0006】

また、 d 個の文書からなる文書の群に対しては、次のような $t \times d$ の単語－文書行列 X が定義できる。

【0007】

$$X = [x_1, x_2, \dots, x_d]$$

ここで、 t 次元のベクトル $x_j = [x_{j1}, x_{j2}, \dots, x_{jt}]'$ は文書 D_j の特徴量ベクトルを表し、記号' は転置を示す。

【0008】

図1は、文書データベースに登録された文書の一例を示す図である。また、図2は、図1に示された文書に出現する漢字の単語を索引語とした単語－文書行列の一例を示す図である。図2において、文書1～3の全てに出現している文字列「について教えて下さい」の中に含まれる「教」の文字は索引語の対象から外されている。図3は、ユーザから実際に入力される質問の一例を示す図である。この質問を図2の索引語を用いて表すと、図4に示す文書－単語行列で表現できる。

【0009】

一般的に、ベクトル空間法を用いた場合、2つの文書 D_r と D_s の類似度 $\text{sim}(D_r, D_s)$ は、以下ようになる。

【0010】

【数2】

$$\text{sim}(D_r, D_s) = \frac{\sum_{i=1}^t x_{ir} x_{is}}{\sqrt{\sum_{i=1}^t x_{ir}^2 \sum_{i=1}^t x_{is}^2}} \quad \text{式(2)}$$

【0011】

図3の質問の意味を基に、この質問と図1の各文書との類似度を判断した場合、図3の質問は図1の文書3に一番類似すると考えられる。しかし、図2及び図4のような特徴量ベクトルを用いると、図1における各文書と図3の質問の類似度は、それぞれ、 $\text{sim}(\text{文書1}, \text{質問}) = 0.5477$ 、 $\text{sim}(\text{文書2}, \text{質問}) = 0.5477$ 、 $\text{sim}(\text{文書3}, \text{質問}) = 0.5477$ となり、全ての文書に対して同じ類似度になってしまう。

【 0 0 1 2 】

このような問題点を解決する手法として「Journal of the American Society for Information Science」（1990年発行）の第41巻第6号第391項から第407項までの記載において提案された、単語の共起に基づいた分析方法（Latent Semantic Analysis; LSA）は、文書のもつ潜在的意味を抽出でき、かつ検索能率が圧倒的に優れている。ここにいう「単語の共起」とは、同一の文書／文に複数の単語が同時に出現することをいう。

【 0 0 1 3 】

LSAは、単語の共起の頻度を示す単語－文書行列を特異値分解（Singular Value Decomposition; SVD）することにより、文書の潜在的意味構造を抽出するものである。得られた特徴量の空間において、互いに関連した文書や単語は近接するように構成される。「Behavior Research Methods, Instruments, & Computers」（1991年発行）の第23巻第2号第229項から第236項までに掲載された論文では、LSAを使用した検索は、ベクトル空間法に比べ、30%効率が良いという結果を報告している。以下、LSAについて具体的に説明する。

【 0 0 1 4 】

LSAでは、まず $t \times d$ の単語－文書行列 X を以下のように特異値分解する。

【 0 0 1 5 】

【数3】

$$X = T_0 S_0 D_0' \quad \text{式(3)}$$

【 0 0 1 6 】

ここで、 T_0 は $t \times m$ の直交行列を表す。 S_0 は m 個の特異値を対角要素とし、かつ対角要素以外はすべて0である $m \times m$ の正方対角行列を表す。 D_0' は $m \times d$ の直交行列を表す。また、 $0 \leq d \leq t$ とし、 S_0 の対角要素は値の大きい順に並んでいるものとする。

【 0 0 1 7 】

更に、LSAでは文書 D_q の $t \times 1$ の特徴量ベクトル x_q に対して次のような変換を行い、 $n \times 1$ のLSA特徴量ベクトル y_q を計算する。

【 0 0 1 8 】

【数 4】

$$y_q = S^{-1} T' x_q \quad \text{式 (4)}$$

【 0 0 1 9 】

ここで、 S は S_0 の対角要素の1番目から n 番目までをとった $n \times n$ の正方対角行列、 T は T_0 の1列目から n 列目まで抜き出した $t \times n$ の行列である。

【 0 0 2 0 】

例として、図2の単語－文書行列に対して特異値分解を行った結果を以下に示す。行列 T_0 、 S_0 、 D_0 はそれぞれ以下ようになる。

【 0 0 2 1 】

【数 5】

$$T_0 = \begin{bmatrix} 0.1787 & -0.3162 & 0.3393 \\ 0.1787 & -0.3162 & 0.3393 \\ 0.1787 & -0.3162 & 0.3393 \\ 0.4314 & -0.3162 & -0.1405 \\ 0.4314 & -0.3162 & -0.1405 \\ 0.1787 & 0.3162 & 0.3393 \\ 0.1787 & 0.3162 & 0.3393 \\ 0.4314 & 0.3162 & -0.1405 \\ 0.4314 & 0.3162 & -0.1405 \\ 0.1787 & 0.3162 & 0.3393 \\ 0.2527 & 0.0000 & -0.4798 \end{bmatrix}$$

【 0 0 2 2 】

【数 6】

$$S_0 = \begin{bmatrix} 2.7979 & 0 & 0 \\ 0 & 2.2361 & 0 \\ 0 & 0 & 1.4736 \end{bmatrix}$$

【 0 0 2 3 】

【数 7】

$$D_0 = \begin{bmatrix} 0.5000 & -0.7071 & 0.5000 \\ 0.5000 & 0.7071 & 0.5000 \\ 0.7071 & 0.0000 & -0.7071 \end{bmatrix}$$

【0 0 2 4】

L S A 特徴量ベクトルの次元 t を 2 とし、図 2 の単語－文書行列の各特徴量ベクトルに対して式 (4) を適用すると、文書 1、2 及び 3 の L S A 特徴ベクトルはそれぞれ $[0.5000, -0.7071]'$ 、 $[0.5000, 0.7071]'$ 、 $[0.7071, 0.0000]'$ となる。また、図 4 の特徴ベクトルに対して式 (4) を適用すると、ユーザの質問の L S A 特徴量ベクトルは $[0.6542, 0]'$ となる。

【0 0 2 5】

上記のように得られた L S A 特徴量ベクトルに対して式 (2) を適用し、図 3 の質問と図 1 に示した各文書との類似度を求めると、図 1 における各文書と図 3 質問の類似度は、それぞれ、 $\text{sim}(\text{文書 1}, \text{質問}) = 0.5774$ 、 $\text{sim}(\text{文書 2}, \text{質問}) = 0.5774$ 、 $\text{sim}(\text{文書 3}, \text{質問}) = 1.0000$ となり、文書 3 が質問と一番類似するという結果が得られる。ネットワークを利用したヘルプシステムの応用などを想定する場合、図 3 の質問をしたユーザに対しては文書データベースに登録された文書 3 の回答文が返信されることになる。

【0 0 2 6】

特異値分解法は、一般的に The Johns Hopkins University Press 社が 1996 年に出版した「Matrix Computations」の第 455 項から第 457 項までの記載において提案されたアルゴリズムがよく用いられる。前記の「Journal of the American Society for Information Science」の論文によると、正方行列 S の行数（または列数） n の値は 50～150 程度にすると良いとの記載がある。また、前記の「Behavior Research Methods, Instruments, & Computers」の論文において、L S A を行う前に特徴ベクトルの各要素を単に 0 または 1 の値をとると定義せずに、上記の出現頻度や文書頻度を用いて前処理するとより効果的であるという結果が報告されている。

【 0 0 2 7 】

【発明が解決しようとする課題】

しかし、上述の文献に提案されている特異値分解法のアルゴリズムでは、与えられた単語－文書行列から特徴量の空間を張る基底ベクトルを計算する過程において行列のバイダイアゴナリゼーション (bidiagonalization) のために $t \times t$ の行列を利用するので、最低でも索引語数 t の二乗 t^2 のオーダーのメモリ空間を必要とする。従って、従来の技術は、膨大な単語数又はデータ数を抱える文書データベースには適用できず、またデータ数の大小に関係なく行列の複雑な演算が必要であるという問題点があった。

【 0 0 2 8 】

本発明はこのような問題点に鑑みてなされたものであり、その目的とするところは、演算処理の容易化および当該演算処理に必要なメモリ容量の低減を図り、効率的に特徴量を抽出するテキストマイニングにおける文書の特徴量抽方法及びその装置を提供することにある。

【 0 0 2 9 】

【課題を解決するための手段】

本発明は、このような目的を達成するため、請求項 1 に記載の発明は、文書の内容を代表する索引語に対応するベクトルからなる単語－文書行列を用いて前記文書の特徴量を抽出するテキストマイニングにおける文書の特徴量抽出方法であって、前記単語－文書行列の各要素には前記索引語に対する寄与分が作用し、コストを最小化する最急降下法に基いて互いに関連した文書および単語が近接する前記特徴量の空間を張る基底ベクトルを計算する基底ベクトル計算ステップと、前記単語－文書行列及び前記基底ベクトルを用いて前記特徴量を正規化するためのパラメータを計算し、該パラメータに基き前記特徴量を抽出する特徴量抽出ステップと、前記単語－文書行列を更新して前記基底ベクトルを適用しない前記単語－文書行列と適用した前記単語－文書行列との差分にする単語－文書行列更新ステップとを備えたことを特徴とする。

【 0 0 3 0 】

また、請求項 2 に記載の発明は、請求項 1 に記載のテキストマイニングにおけ

る文書の特徴量抽出方法において、前記コストは、前記基底ベクトルを適用しない前記単語一文書行列と適用した前記単語一文書行列との差分の二次コストとして定義されることを特徴とする。

【 0 0 3 1 】

また、請求項 3 に記載の発明は、請求項 1 または 2 に記載のテキストマイニングにおける文書の特徴量抽出方法において、前記基底ベクトル計算ステップは、前記基底ベクトルの値を初期化する初期化ステップと、前記基底ベクトルの値を更新する基底ベクトル更新ステップと、前記基底ベクトルの値の変化度合いを求める変化度合い計算ステップと、前記基底ベクトルの値の変化度合いを用いて繰り返し処理を終了するかどうかを判別する判別ステップと、前記繰り返し処理の回数を数える計数ステップとを備えたことを特徴とする。

【 0 0 3 2 】

また、請求項 4 に記載の発明は、請求項 3 に記載のテキストマイニングにおける文書の特徴量抽出方法において、前記基底ベクトル更新ステップは、前記基底ベクトルの現在値と、前記単語一文書行列と、前記基底ベクトルの更新度合いを制御する更新率とを用いて前記基底ベクトルを更新することを特徴とする。

【 0 0 3 3 】

また、請求項 5 に記載の発明は、請求項 1 ～ 4 のいずれか 1 項に記載のテキストマイニングにおける文書の特徴量抽出方法において、前記特徴量の抽出に必要とされる全ての前記基底ベクトル及び前記正規化パラメータを既に取得している場合は、前記基底ベクトル計算ステップ及び前記特徴量抽出ステップにおける前記正規化パラメータの計算を省略し、前記特徴量抽出ステップは、既に取得している前記基底ベクトル及び前記正規化パラメータを用いて前記特徴量を抽出することを特徴とする。

【 0 0 3 4 】

また、請求項 6 に記載の発明は、テキストマイニングにおける文書の特徴量抽出装置において、文書の内容を代表する索引語に対応するベクトルからなる単語一文書行列を用いて前記文書の特徴量を抽出するテキストマイニングにおける文書の特徴量抽出装置であって、前記単語一文書行列の各要素には前記索引語に対

する寄与分が作用し、コストを最小化する最急降下法に基いて互に関連した文書および単語が近接する前記特徴量の空間を張る基底ベクトルを計算する基底ベクトル計算手段と、前記単語一文書行列及び前記基底ベクトルを用いて前記特徴量を正規化するためのパラメータを計算し、該パラメータに基き前記特徴量を抽出する特徴量抽出手段と、前記単語一文書行列を更新して前記基底ベクトルを適用しない前記単語一文書行列と適用した前記単語一文書行列との差分にする単語一文書行列更新手段とを備えたことを特徴とする。

【 0 0 3 5 】

また、請求項 7 に記載の発明は、請求項 6 に記載のテキストマイニングにおける文書の特徴量抽出装置において、前記コストは、前記基底ベクトルを適用しない前記単語一文書行列と適用した前記単語一文書行列との差分の二次コストとして定義されることを特徴とする。

【 0 0 3 6 】

また、請求項 8 に記載の発明は、請求項 6 または 7 に記載のテキストマイニングにおける文書の特徴量抽出装置において、前記基底ベクトル計算手段は、前記基底ベクトルの値を初期化する初期化手段と、前記基底ベクトルの値を更新する基底ベクトル更新手段と、前記基底ベクトルの値の変化度合いを求める変化度合い計算手段と、前記基底ベクトルの値の変化度合いを用いて繰り返し処理を終了するかどうかを判別する判別手段と、前記繰り返し処理の回数を数える計数手段とを備えたことを特徴とする。

【 0 0 3 7 】

また、請求項 9 に記載の発明は、請求項 8 に記載のテキストマイニングにおける文書の特徴量抽出装置において、前記基底ベクトル更新手段は、前記基底ベクトルの現在値と、前記単語一文書行列と、前記基底ベクトルの更新度合いを制御する更新率とを用いて前記基底ベクトルを更新することを特徴とする。

【 0 0 3 8 】

また、請求項 1 0 に記載の発明は、請求項 6 ～ 9 のいずれか 1 項に記載のテキストマイニングにおける文書の特徴量抽出装置において、前記特徴量の抽出に必要とされる全ての前記基底ベクトル及び前記正規化パラメータを既に取得してい

る場合は、前記基底ベクトル計算手段及び前記特徴量抽出手段における前記正規化パラメータの計算を省略し、前記特徴量抽出手段は、既を取得している前記基底ベクトル及び前記正規化パラメータを用いて前記特徴量を抽出することを特徴とする。

【 0 0 3 9 】

更に、請求項 1 1 に記載の発明は、文書の内容を代表する索引語に対応するベクトルからなる単語－文書行列を用いて前記文書の特徴量を抽出するテキストマイニングにおける文書の特徴量抽出装置において実行される特徴量抽出プログラム・プロダクトであって、前記単語－文書行列の各要素には前記索引語に対する寄与分が作用し、コストを最小化する最急降下法に基いて互いに関連した文書および単語が近接する前記特徴量の空間を張る基底ベクトルを計算する基底ベクトル計算ステップと、前記単語－文書行列及び前記基底ベクトルを用いて前記特徴量を正規化するためのパラメータを計算し、該パラメータに基き前記特徴量を抽出する特徴量抽出ステップと、前記単語－文書行列を更新して前記基底ベクトルを適用しない前記単語－文書行列と適用した前記単語－文書行列との差分にする単語－文書行列更新ステップとを備えたことを特徴とする。

【 0 0 4 0 】

本明細書によって開示される特徴量抽出装置は、以下の手段によって構成される。即ち、元の単語－文書行列と基底ベクトルを適用した単語－文書行列との差分の二次関数をコストとして定義し、そのコストに対して最急降下法を適用して基底ベクトルを計算する基底ベクトル計算手段と、単語－文書行列及び基底ベクトルを用いて、特徴量を正規化するためのパラメータを計算し、各文書に対して特徴量を抽出する特徴量抽出手段と、特徴量抽出手段の実行間で重複した特徴量を抽出しないように上記の差分で単語－文書行列を更新する単語－文書行列更新手段と、上記各手段の実行を制御する特徴量抽出制御手段とを備えていれば足りる。

【 0 0 4 1 】

基底ベクトル計算手段は、入力された単語－文書行列を基に計算を繰り返し、最終的に 1 つの基底ベクトルを算出する。繰り返しの処理は、各繰り返し処理間

で基底ベクトルの変化度合いが所定の基準値以下になったときに終了する。特徴量抽出手段は、入力された基底ベクトル及び単語－文書行列を基に、特徴量を正規化するためのパラメータを計算し、各文書に対して1つの特徴量を抽出する。単語－文書行列更新手段は、入力された基底ベクトルを基に、単語－文書行列を更新する。

【 0 0 4 2 】

特徴量抽出制御手段は、基底ベクトル計算手段、特徴量抽出手段、及び単語－文書行列更新手段を制御し、ユーザーにより定義された特徴量の数を満たすまで、各手段の実行を繰り返す。但し、基底ベクトル及び正規化パラメータが既に計算されている場合には、基底ベクトル計算手段の実行及び特徴量抽出手段における正規化パラメータの計算を省略される。そして、既に取得している基底ベクトル及び正規化パラメータを組み込んだ構成で特徴量抽出を行うことになる。

【 0 0 4 3 】

【発明の実施の形態】

図5は、本発明に係る特徴量抽出装置の一実施例を示す図である。図5に示すように、特徴量抽出制御手段200は、単語－文書行列更新手段210と、基底ベクトル計算手段220と、特徴量抽出手段230とを備える。100は単語－文書行列データファイル、300は基底ベクトルデータファイル、400は特徴量データファイル、450は正規化パラメータデータファイルである。単語－文書行列データファイル100には、収集された文書データの単語－文書行列が記憶されている。単語－文書行列更新手段210は第1回目の繰り返し処理で単語－文書行列データファイル100から単語－文書行列を読み込み、その単語－文書行列を更新せずに基底ベクトル計算手段220及び特徴量抽出手段230に渡す。

【 0 0 4 4 】

第2回目の繰り返し処理以降では、基底ベクトル計算手段220から渡された基底ベクトルを基に単語－文書行列を更新し、その結果を基底ベクトル計算手段220及び特徴量抽出手段230に渡す。基底ベクトル計算手段220は、単語－文書行列更新手段210から渡された単語－文書行列を基に繰り返し処理によ

り1つの基底ベクトルを計算する。そして、各繰り返し処理で基底ベクトルの変化度合いを監視し、変化度合いが所定の基準値以下になったときに繰り返しの処理を終了する。基底ベクトル計算手段220は、計算した基底ベクトルを基底ベクトルデータファイル300に格納すると同時に、単語一文書行列更新手段210及び特徴量抽出手段230に渡す。特徴量抽出手段230は単語一文書行列更新手段210から渡された単語一文書行列及び基底ベクトル計算手段220から渡された基底ベクトルを基に各文書に対して1つの特徴量を抽出する。その結果を特徴量データファイル400に格納すると同時に、それらの特徴量を正規化するためのパラメータを正規化パラメータデータファイル450に記録する。

【0045】

単語一文書行列更新手段210、基底ベクトル計算手段220及び特徴量抽出手段230による、上述の実行を1回の繰り返しとする。繰り返し処理の回数を添字*i*で、ユーザーが指定した特徴量の数を添字*n*で示す。特徴量抽出制御手段200では、 $i = n$ の条件を満たすまで、処理を一単位ずつ繰り返す。また、必要とされる全ての基底ベクトル及び正規化パラメータを既に取得しており、これらの値が既知の場合は、基底ベクトル計算手段220の実行及び特徴量抽出手段230における正規化パラメータの計算を省略し、既知の基底ベクトル及び正規化パラメータを組み込んだ単語一文書行列更新手段210及び特徴量抽出手段230のみで特徴量抽出制御手段200を構成する。

【0046】

図6は、本発明を実施するハードウェア構成の一例を示す図である。図6に示すように、特徴量抽出装置は、装置全体の制御を行う中央処理装置（Central Processor Unit; CPU）10と、プログラムが格納され又はプログラムの実行に必要な一時データ格納領域を提供するメモリ20と、データを入力するためのキーボード30と、表示画面を生成するディスプレイ40とを備える。単語一文書行列データファイル100、基底ベクトルデータファイル300、特徴量データファイル400、正規化パラメータデータファイル450及び特徴量抽出制御手段200によって実行されるプログラムはメモリ20に格納されている。

【0047】

このような構成をとることにより、キーボード30又はディスプレイ40上の所定の位置指定するマウス等によりユーザーの指示を受けたCPU10によって特徴量抽出が行われることとなる。なお、図5に示す例では、特徴量抽出制御手段200はスタンドアロンの構成としているが、他のシステムに組み込んだ構成とすることも可能であることは言うまでもない。

【0048】

図7は、単語一文書行列データファイルの構成図である。図7において、101-1, 101-2, ..., 101-dはd個からなるt次元の単語一文書データに対応する。ここで、 $X = [x_1, x_2, \dots, x_d]$ 、 $x_j = [x_{j1}, x_{j2}, \dots, x_{jt}]'$ を定義し、単語一文書データ101を $t \times d$ の行列Xで示す。

【0049】

図8は、計算された基底ベクトルが格納された基底ベクトルデータファイルの構成図である。図8において、301-1, 301-2, ..., 301-nはn個からなるt次元の基底ベクトルデータに対応する。i番目の要素301-iは、図5におけるi回目の繰り返し処理における基底ベクトル計算手段220の出力値に対応する。以下の説明では、この要素を $t \times 1$ の列ベクトル $w_i = [w_{i1}, w_{i2}, \dots, w_{it}]'$ で示す。

【0050】

図9は、特徴量データファイルの構成図である。図9において、401-1, 401-2, ..., 401-nはn個からなるd次元の特徴量データに対応する。i番目の要素401-iは図5におけるi回目の繰り返し処理における特徴量抽出手段230による特徴量の出力値に対応する。この要素を $1 \times d$ の行ベクトル $y_i = [y_{i1}, y_{i2}, \dots, y_{id}]$ で示す。

【0051】

図10は、正規化パラメータデータファイルの構成図である。図10において、451-1, 452-2, ..., 451-nはn個からなる正規化パラメータデータに対応する。i番目の要素451-iは図5におけるi回目の繰り返し処理での特徴量抽出手段230による正規化パラメータの出力値に対応する。この要素を p_i で示す。

【0052】

以上の諸定義を使用し、本実施形態に係る特徴量抽出の実現方式を詳細に説明する。単語一文書行列更新手段210では、 $i = 1$ の場合、即ち繰り返し処理の1回目の実行に限り、 X を単語一文書行列データファイル100から読み込み、何ら演算を行うことなく $t \times d$ の行列 E に格納する。従って、 $E = [e_1, e_2, \dots, e_d]$ 、 $e_j = [e_{j1}, e_{j2}, \dots, e_{jt}]' = [x_{j1}, x_{j2}, \dots, x_{jt}]'$ となる。前の繰り返し処理で抽出された特徴量が重複して抽出されないために、図5における i 回目の繰り返しで下のように E をその現在値及び1つ前の繰り返し処理において計算された基底ベクトルを用いて更新し、その結果を基底ベクトル計算手段220に渡す。この処理によって格納される、 E の i 番目の処理結果 $E(i)$ は、式(5)のように表される。

【0053】

【数8】

$$E(i) = \begin{cases} X, & \text{for } i = 1 \\ E(i-1) - w_{i-1}(w'_{i-1}E(i-1)) & , \quad \text{otherwise} \end{cases} \quad \text{式(5)}$$

【0054】

ここで、 $E(i) = [e_1(i), e_2(i), \dots, e_d(i)]$ であり、 $E(i)$ の各要素 $e_j(i)$ は $e_j(i) = [e_{j1}(i), e_{j2}(i), \dots, e_{jt}(i)]'$ で定義される。即ち、 $i \geq 2$ の場合は、単語一文書行列は基底ベクトルを適用しない単語一文書行列から基底ベクトルを適用した単語一文書行列を引いた差分に更新される。

【0055】

図11は、基底ベクトル計算手段における基底ベクトルの計算の流れ図である。図11における k 回目の繰り返しでの w_i の値を $w_i(k) = [w_{i1}(k), w_{i2}(k), \dots, w_{it}(k)]'$ で示す。まず、ステップS500にて添字 k を1で初期化する。続いてステップS510へ移行し、 $w_i(1)$ の各要素を $-C$ から C までの間の任意の値で初期化する。ここで、 C の値は正の小さい数であり、例えば $C = 0.01$ としてもよい。ステップS520では、互いに関連した文書

や単語が近接する特徴量の空間を張る基底ベクトルを計算するため、式(6)に示す二次コストを設ける。

【0056】

【数9】

$$\frac{1}{2d} \sum_{m=1}^d \sum_{i=1}^t (e_{im}(i) - w_{1i} \tilde{y}_{im})^2 \quad \text{式(6)}$$

【0057】

ここで、「単語が近接する」とは、特徴量の空間の中で複数の単語の位置が近接することを言い、「文書が近接する」とは、複数の文書の各々に含まれる単語の位置が特徴量の空間の中で近接することを言う。また、コストとは最小化したい対象を言い、本実施形態で定義されるコストは式(6)のように基底ベクトルを適用しない単語一文書行列と基底ベクトルを適用した単語一文書行列との差分の二次関数として定義される。ここで、

【0058】

【外1】

\tilde{y}_{im}

【0059】

は次のように定義される $1 \times d$ のベクトル

【0060】

【外2】

\tilde{y}_i

【0061】

の m 番目の要素である。

【0062】

【数10】

$$\tilde{y}_i = [\tilde{y}_{i1}, \tilde{y}_{i2}, \dots, \tilde{y}_{id}] = w_i' E(i) \quad \text{式(7)}$$

【0063】

上記のコストに対して最急降下法を適用して w_i の値を式(8)のように更新する。

【 0 0 6 4 】

【 数 1 1 】

$$w_i(k+1) = w_i(k) + \frac{\mu_i(k)}{d} (E(i) - w_i(k) z_i(k)) z_i(k)' \quad \text{式 (8)}$$

【 0 0 6 5 】

ここで、 $\mu_i(k)$ は k 回目の繰返しでの更新の度合いを制御する更新率で、 k が 1 のときに正の小さい数で初期化し、例えば $\mu_i(1) = 0.1$ としてもよい。 k が加算される度に徐々に値を減少させ、あるいは k の値によらず一定値とすることも可能である。また、 $z_i(k)$ は次のように定義される。

【 0 0 6 6 】

【 数 1 2 】

$$z_i(k) = w_i(k)' E(i) \quad \text{式 (9)}$$

【 0 0 6 7 】

ステップ S 5 3 0 では次のように w_i の変化度合いを示す δ_i を求める。

【 0 0 6 8 】

【 数 1 3 】

$$\delta_i(k) = \sqrt{\sum_{j=1}^t (w_{ji}(k+1) - w_{ji}(k))^2} \quad \text{式 (10)}$$

【 0 0 6 9 】

ステップ S 5 4 0 では $\delta_i(k)$ の値を基に処理を終了するかどうかを判別する。判別の結果、終了すると判断した場合はステップ S 5 6 0 へ進み、そうでない場合はステップ S 5 5 0 へ進む。ここで、図 1 1 における β_i は正の小さい数であり、例えば $\beta_i = 1 \times 10^{-6}$ とすることができる。

【 0 0 7 0 】

ステップ S 5 5 0 ではカウンタ k の値を 1 つ増やし、ステップ S 5 2 0 に戻る。ステップ S 5 6 0 では w_i を基底ベクトルデータファイル 3 0 0 に i 番目のデータとして格納すると同時に、単語一文書行列更新手段 2 1 0 及び特徴量抽出手段 2 3 0 に渡す。特徴量抽出手段 2 3 0 では次のように特徴量 y_i 及び正規化パ

ラメータ p_i を計算して、その結果それぞれを特徴量データファイル 4 0 0 及び正規化パラメータデータファイル 4 5 0 に i 番目のデータとして格納する。

【 0 0 7 1 】

【数 1 4】

$$y_i = \tilde{y}_i / p_i, \quad \text{式(11)}$$

【 0 0 7 2 】

ここで、 p_i は次のように定義される。

【 0 0 7 3 】

【数 1 5】

$$p_i = \sqrt{\sum_{j=1}^d \tilde{y}_{ij}^2} \quad \text{式 (12)}$$

【 0 0 7 4 】

図 1. 2 は、本実施形態に係る特徴量抽出装置を用いた文書自動分類装置の一例を示す図である。図 1. 2 において、6 0 1 は単語－文書行列計算手段、6 0 2 は分類手段である。分類手段 6 0 2 は、1 9 9 3 年に発行された「Journal of Intelligent and Fuzzy Systems」の第 1 巻第 1 号第 1 項から第 2 5 項で述べられている方法により行うことができる。

【 0 0 7 5 】

文書データベースに格納されている文書データは、文書自動分類装置に取り込まれる。文書自動分類装置では、単語－文書行列計算手段 6 0 1 で単語－文書行列の計算を行い、その結果を特徴量抽出制御手段 2 0 0 に渡す。特徴量抽出制御手段 2 0 0 は受け取った単語－文書行列から特徴量を抽出し、その結果を分類手段 6 0 2 に出力する。分類手段 6 0 2 では、入力された特徴量を基に分類の結果が出力される。

【 0 0 7 6 】

本発明の評価を、図 1 の文書や図 3 の質問のような文書データを含んだ、入試制度に関する実際の文書データにより特徴量抽出の評価を行った。本発明は、従来の L S A を使用した場合と同じ性質の特徴量を抽出することが確認できた。

【 0 0 7 7 】

次に、使用されるメモリ空間のサイズに関して、単語数 t が文書数 d よりかなり大きい ($t \gg d$) といった実際の場合において、従来の LSA が少なくとも、 t^2 のオーダーを必要するのに対して、本発明は各々基底ベクトルの計算のために高々 $t \cdot d$ のオーダーのメモリサイズで足りる。また、従来の技術を実現するには、複雑な行列演算装置が必要であるが、本方式は四則演算程度を行う装置があれば容易に実現することができる。即ち、本発明によれば、LSA による特徴量抽出と同等の効果を、より小さいメモリ空間、より簡単なプログラムにより得ることができる。また、この簡単なプログラムは DSP (Digital Signal Processor) におとすことができるため、特徴量抽出専用のチップを簡単に作成することが出来る。

【0078】

以下、図1の文書及び図3の質問に対して本実施形態に係る特徴量抽出装置を実行した各手段の結果を示す。

【0079】

A. 図1の文書

まず、図2の単語－文書行列を X とする。

【0080】

I. 特徴量抽出制御手段200における第1回目の繰り返し ($i = 1$)

単語－文書行列更新手段210では式(5)より

【0081】

【数16】

$$E(1) = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

【 0 0 8 2 】

を基底ベクトル計算手段 2 2 0 及び特徴量抽出手段 2 3 0 へ出力する。

【 0 0 8 3 】

基底ベクトル計算手段 2 2 0 では、 $w_1(1)$ を

$[0.0100, -0.0100, 0.0100, -0.0100, 0.0100, -0.0100, 0.0100, -0.0100, 0.0100, -0.0100, 0.0100, 0.0100]'$

で、 μ_1 を固定の値 0.1 で、 β_1 を 1×10^{-6} で初期化し、以下のように図 1 1 の計算を 1 3 2 回繰り返した後、基底ベクトル $w_1 =$

$[0.1787, 0.1787, 0.1787, 0.4314, 0.4314, 0.1787, 0.1787, 0.4314, 0.4314, 0.1787, 0.2527]'$

を基底ベクトルデータファイル 3 0 0 に格納すると共に、特徴量抽出手段 2 3 0 、及び単語一文書行列更新手段 2 1 0 へ出力する。

【 0 0 8 4 】

* 基底ベクトル計算手段 2 2 0 における第 1 回目の繰り返し ($k = 1$)

式 (8) より、

$w_1(2) =$

$[0.0103, -0.0097, 0.0103, -0.0093, 0.0107, -0.0103, 0.0097, -0.0100, 0.0100, -0.0103, 0.0103]'$

$w_1(2) - w_1(1) = 10^{-3} \times$

$[0.3332, 0.3334, 0.3332, 0.6668, 0.6666, -0.3332, -0.3334, 0.0001, -0.0001, -0.3332, 0.3332]'$

$\delta_1(1) = 0.0013$

* 基底ベクトル計算手段 2 2 0 における第 2 回目の繰り返し ($k = 2$)

式 (8) より、

$w_1(3) =$

$[0.0107, -0.0093, 0.0107, -0.0085, 0.0115, -0.0107, 0.0093, -0.0100, 0.0100, -0.0107, 0.0107]'$

$w_1(3) - w_1(2) = 10^{-3} \times$

$[0.4110, 0.4112, 0.4110, 0.8001, 0.7998, -0.3665, -0.3668, 0.0224,$

0.0221, -0.3665, 0.3887]'

$$\delta_1(2) = 0.0015$$

中 略

* 基底ベクトル計算手段 2 2 0 における第 1 3 2 回目の繰り返し ($k = 1 3 2$)

式 (8) より、

$$w_1(1 3 3) =$$

[0.1787, 0.1787, 0.1787, 0.4314, 0.4314, 0.1787, 0.1787, 0.4314, 0.4314, 0.1787, 0.2527]'

$$w_1(1 3 3) - w_1(1 3 2) = 1 0^{-6} \times$$

[-0.3020, -0.3020, -0.3020, -0.3020, -0.3020, 0.3020, 0.3020, 0.3020, 0.3020, 0.3020, 0.0000]'

$$\delta_1(1 3 2) = 9.5500 \times 1 0^{-7}$$

特徴量抽出手段 2 3 0 では式 (1 1) 及び式 (1 2) の演算を行い、

$$y_1 = [0.5000, 0.5000, 0.7071]$$

及び

$$p_1 = 2.7979$$

をそれぞれ特徴量データファイル 4 0 0 及び正規化パラメータデータファイル 4 5 0 へ出力する。

【 0 0 8 5 】

I I . 特徴量抽出制御手段 2 0 0 における第 2 回目の繰り返し ($i = 2$)

単語一文書行列更新手段 2 1 0 では式 (5) より

【 0 0 8 6 】

【数 1 7】

$$E(2) = \begin{bmatrix} 0.7500 & -0.2500 & -0.3536 \\ 0.7500 & -0.2500 & -0.3536 \\ 0.7500 & -0.2500 & -0.3536 \\ 0.3964 & -0.6036 & 0.1464 \\ 0.3964 & -0.6036 & 0.1464 \\ -0.2500 & 0.7500 & -0.3535 \\ -0.2500 & 0.7500 & -0.3535 \\ -0.6036 & 0.3965 & 0.1465 \\ -0.6036 & 0.3965 & 0.1465 \\ -0.2500 & 0.7500 & -0.3535 \\ -0.3536 & -0.3535 & 0.5000 \end{bmatrix}$$

【 0 0 8 7】

を基底ベクトル計算手段 2 2 0 及び特徴量抽出手段 2 3 0 へ出力する。

【 0 0 8 8】

基底ベクトル計算手段 2 2 0 では、 $w_2(1)$ を

$$[0.0100, -0.0100, 0.0100, -0.0100, 0.0100, -0.0100, 0.0100, -0.0100, 0.0100, -0.0100, 0.0100]'$$

で、 μ_2 を固定の値 0.1 で、 β_2 を 1×10^{-6} で初期化し、図 1 1 の計算を 1 1 9 回繰り返した後、基底ベクトル $w_2 =$

$$[0.3162, 0.3162, 0.3162, 0.3162, 0.3162, -0.3162, -0.3162, -0.3162, -0.3162, -0.3162, 0.0000]'$$

を基底ベクトルデータファイル 3 0 0 に格納すると共に、特徴量抽出手段 2 3 0 、及び単語一文書行列更新手段 2 1 0 へ出力する。

【 0 0 8 9】

* 基底ベクトル計算手段 2 2 0 における第 1 回目の繰り返し ($k = 1$)

式 (8) より、

$$w_2(2) =$$

$$[0.0102, -0.0098, 0.0102, -0.0096, 0.0104, -0.0105, 0.0095, -0.0103, 0.0097, -0.0105, 0.0102]'$$

$$w_2(2) - w_2(1) = 10^{-3} \times$$

$$[0.2154, 0.2156, 0.2154, 0.3822, 0.3821, -0.4511, -0.4513, -0.2844$$

, -0.2846, -0.4511, 0.1666]'

$$\delta_2(1) = 0.0011$$

* 基底ベクトル計算手段 2 2 0 における第 2 回目の繰り返し (k = 2)

式 (8) より、

$$w_2(3) =$$

[0.0105, -0.0095, 0.0105, -0.0092, 0.0108, -0.0110, 0.0090, -0.0106, 0.0094, -0.0110, 0.0103]'

$$w_2(3) - w_2(2) = 10^{-3} \times$$

[0.2624, 0.2626, 0.2624, 0.4413, 0.4411, -0.5152, -0.5154, -0.3364, -0.3366, -0.5152, 0.1786]'

$$\delta_2(2) = 0.0013$$

中 略

* 基底ベクトル計算手段 2 2 0 における第 1 1 9 回目の繰り返し (k = 1 1 9)

式 (8) より、

$$w_2(120) =$$

[0.3162, 0.3162, 0.3162, 0.3162, 0.3162, -0.3162, -0.3162, -0.3162, -0.3162, 0.0000]'

$$w_2(120) - w_2(119) = 10^{-6} \times$$

[0.3327, 0.3333, 0.3327, -0.1375, -0.1381, 0.3332, 0.3326, -0.1377, -0.1383, 0.3332, -0.4712]'

$$\delta_2(119) = 9.8141 \times 10^{-7}$$

—特徴量抽出手段 2 3 0 では式 (1 1) 及び式 (1 2) の演算を行い、

$$y_2 = [0.7071, -0.7071, -0.0000]$$

及び

$$p_2 = 2.2361$$

をそれぞれ特徴量データファイル 4 0 0 及び正規化パラメータデータファイル 4

50へ出力する。

【0090】

上記の結果から図1における文書1, 2, 3の特徴量ベクトルはそれぞれ $[0.5000, 0.7071]'$ 、 $[0.5000, -0.7071]'$ 、 $[0.7071, -0.0000]'$ となる。これらは、従来例の説明において示された各文書のLSAの特徴量と比較すると、第二番目の要素の符号が逆になっているが同一の絶対値をとる。従って、式(2)の類似度の計算に関してLSAの特徴量と同じ性質を持つ。

【0091】

B. 図3の質問

ここでは、図1の文書の特徴量抽出の際に基底ベクトルデータファイル300に格納された基底ベクトル及び正規化パラメータデータファイル450に格納された正規化パラメータを用いるので、基底ベクトル計算手段220の実行及び特徴量抽出手段における正規化パラメータの計算を省略する。図3の質問をXとする。

【0092】

I. 特徴量抽出手段200における第1回目の繰り返し ($i = 1$)

単語一文書行列更新手段210では、式(5)より

【0093】

【数18】

$$E(1) = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

【0094】

を特徴量抽出手段230へ出力する。

【 0 0 9 5 】

特徴量抽出手段 2 3 0 では、図 1 の文書の特徴量抽出の際に得られた特徴量ベクトル w_1 及び正規化パラメータ p_1 を用いて式 (1 1) 及び式 (1 2) の演算を行い

$$y_1 = [0.6542]$$

を特徴量データファイル 4 0 0 へ出力する。

【 0 0 9 6 】

I I. 特徴量抽出制御手段 2 0 0 における第 2 回目の繰り返し ($i = 2$)

単語一文書行列更新手段 2 1 0 では、図 1 に示す文書の特徴量抽出を行う際に得られた特徴量ベクトル w_1 を用いて、式 (5) より

【 0 0 9 7 】

【 数 1 9 】

$$E(2) = \begin{bmatrix} -0.3271, \\ -0.3271, \\ 0.6729, \\ 0.2103, \\ 0.2103, \\ 0.6729, \\ 0.6729, \\ -0.7897, \\ 0.2103, \\ -0.3271, \\ -0.4626 \end{bmatrix}$$

【 0 0 9 8 】

を特徴量抽出手段 2 3 0 へ出力する。

【 0 0 9 9 】

特徴量抽出手段 2 3 0 では図 1 の文書の特徴量抽出の際に得られた特徴量ベクトル w_2 及び正規化パラメータ p_2 を用いて、式 (1 1) 及び式 (1 2) の演算を行い、

$$y_2 = [-0.0000]$$

を特徴量データファイル 4 0 0 へ出力する。

【 0 1 0 0 】

上記の結果から図3の質問の特徴量ベクトルは $[0.6542, -0.0000]'$ となり、従来例の説明で示した値と比較すると2番目の要素は同一の絶対値をとる。

【0 1 0 1】

【発明の効果】

以上説明したように、本発明によれば、文書の内容を代表する索引語に対応するベクトルからなる単語－文書行列を用いて文書の特徴量を抽出するテキストマイニングにおける文書の特徴量抽出方法であって、単語－文書行列の各要素には前記索引語に対する寄与分が作用し、コストを最小化する最急降下法に基いて互に関連した文書および単語が近接する特徴量の空間を張る基底ベクトルを計算する基底ベクトル計算ステップと、単語－文書行列及び基底ベクトルを用いて特徴量を正規化するためのパラメータを計算し、パラメータに基き特徴量を抽出する特徴量抽出ステップと、単語－文書行列を更新して基底ベクトルを適用しない単語－文書行列と適用した単語－文書行列との差分にする単語－文書行列更新ステップとを備えたので、テキストマイニングにおける文書の特徴量抽出に関し、LSAを実行可能な装置よりも小さいメモリ空間でLSAと同じ性質を持つ特徴量を抽出することができる。また、LSAと同じ性質を持つ特徴量を抽出するための専用ソフトウェアやハードウェアを容易に作成することが可能となる。

【図面の簡単な説明】

【図1】

文書データベースに登録された文書の一例を示す図である。

【図2】

図1に示された文書に出現する漢字の単語を索引語とした単語－文書行列の一例を示す図である。

【図3】

ユーザから実際に入力される質問の一例を示す図である。

【図4】

図3から得られた単語－文書行列を示す図である。

【図5】

本発明に係る特徴量抽出装置の一実施例を示す図である。

【図 6】

本発明を実施するハードウェア構成の一例を示す図である。

【図 7】

単語－文書行列データファイルの構成図である。

【図 8】

計算された基底ベクトルが格納された基底ベクトルデータファイルの構成図である。

【図 9】

特徴量データファイルの構成図である。

【図 1 0】

正規化パラメータデータファイルの構成図である。

【図 1 1】

基底ベクトル計算手段における基底ベクトルの計算の流れ図である。

【図 1 2】

本発明の一実施形態に係る特徴量抽出装置を用いた文書自動分類装置の一例を示す図である。

【符号の説明】

1 0 C P U

2 0 メモリ

3 0 キーボード

4 0 ディスプレイ

1 0 0 単語－文書行列データファイル

1 0 1 - 1、1 0 1 - 2、1 0 1 - 3、1 0 1 - d 単語－文書データ

2 0 0 特徴量抽出制御手段

2 1 0 単語－文書行列更新手段

2 2 0 基底ベクトル計算手段

2 3 0 特徴量抽出手段

3 0 0 基底ベクトルデータファイル

3 0 1 - 1、3 0 1 - 2、3 0 1 - 3、3 0 1 - n 基底ベクトルデータ

4 0 0 特徴量データファイル

4 0 1 - 1、4 0 1 - 2、4 0 1 - 3、4 0 1 - n 特徴量データ

4 5 0 正規化パラメータデータファイル

4 5 1 - 1、4 5 1 - 2、4 5 1 - 3、4 5 1 - n 正規化パラメータデータ

6 0 1 単語一文書行列計算手段

6 0 2 分類手段

【書類名】 図面

【図1】

文書1： 推薦入試と大学の教育方針について教えてください。

文書2： 浪人している学生のための秋季入学の概要について
教えてください。

文書3： 秋季入学と貴学の教育方針について教えてください。

【図2】

<div> <div>文書</div> <div>索引語</div> </div>	文書1	文書2	文書3
推薦	1	0	0
入試	1	0	0
大学	1	0	0
教育	1	0	1
方針	1	0	1
浪人	0	1	0
学生	0	1	0
秋季	0	1	1
入学	0	1	1
概要	0	1	0
貴学	0	0	1

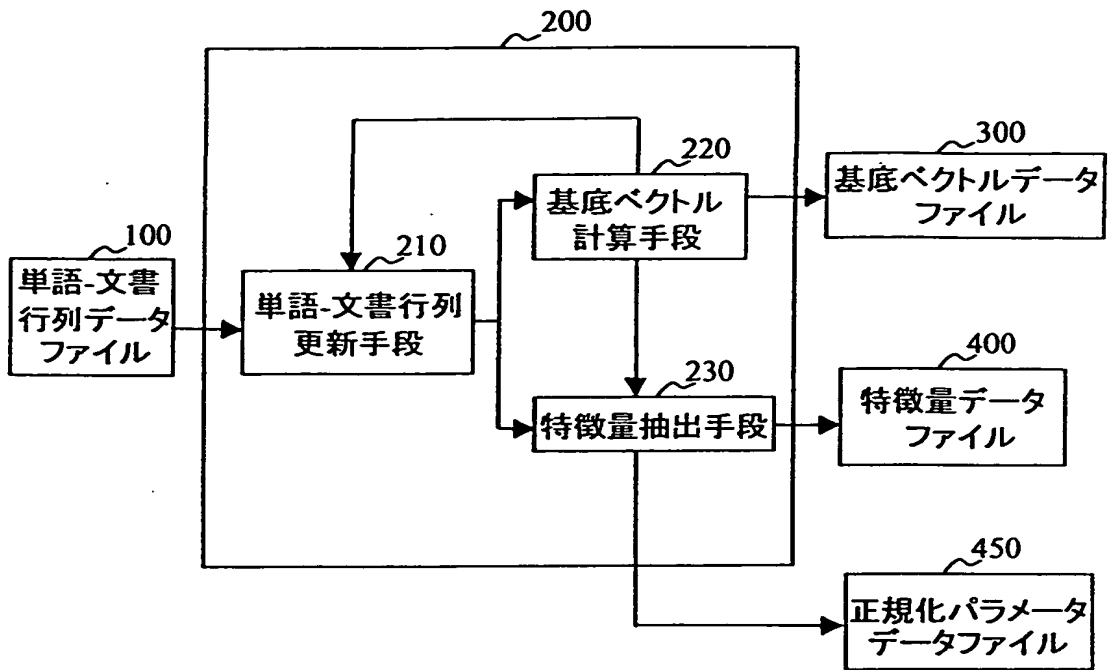
【図3】

質問： 浪人の学生を対象した入学の制度及び大学の教育方針
について教えてください。

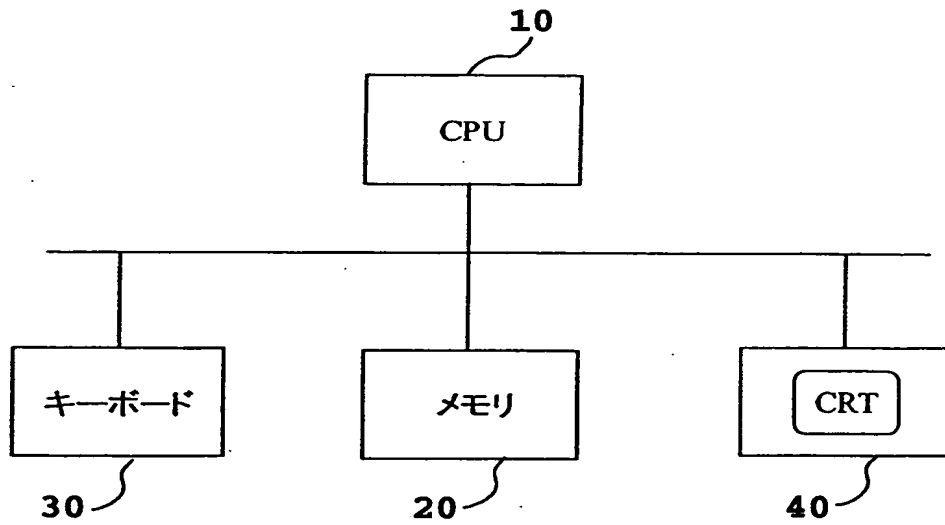
【図 4】

推薦	0
入試	0
大学	1
教育	1
方針	1
浪人	1
学生	1
秋季	0
入学	1
概要	0
貴学	0

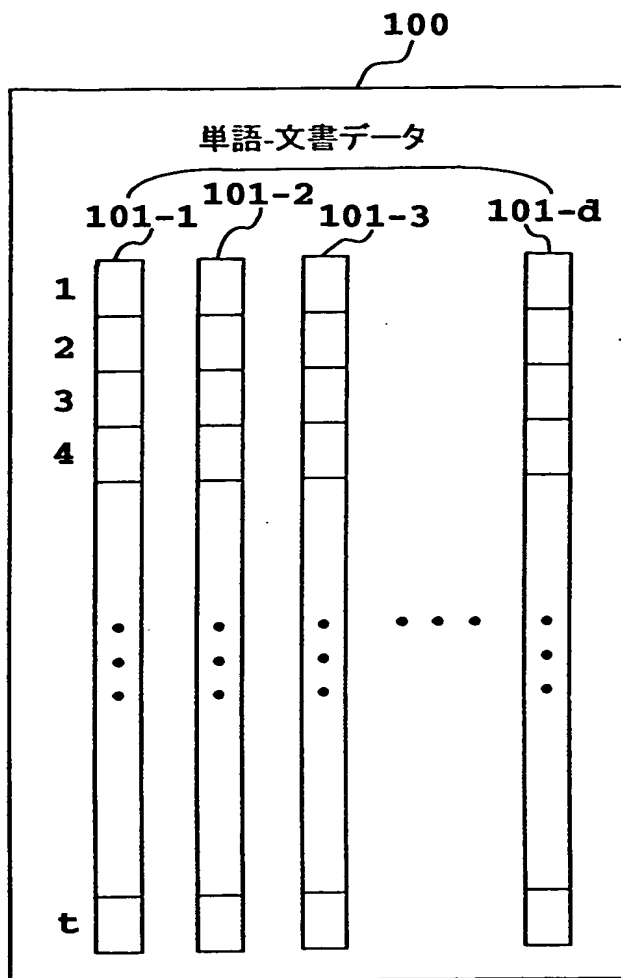
【図 5】



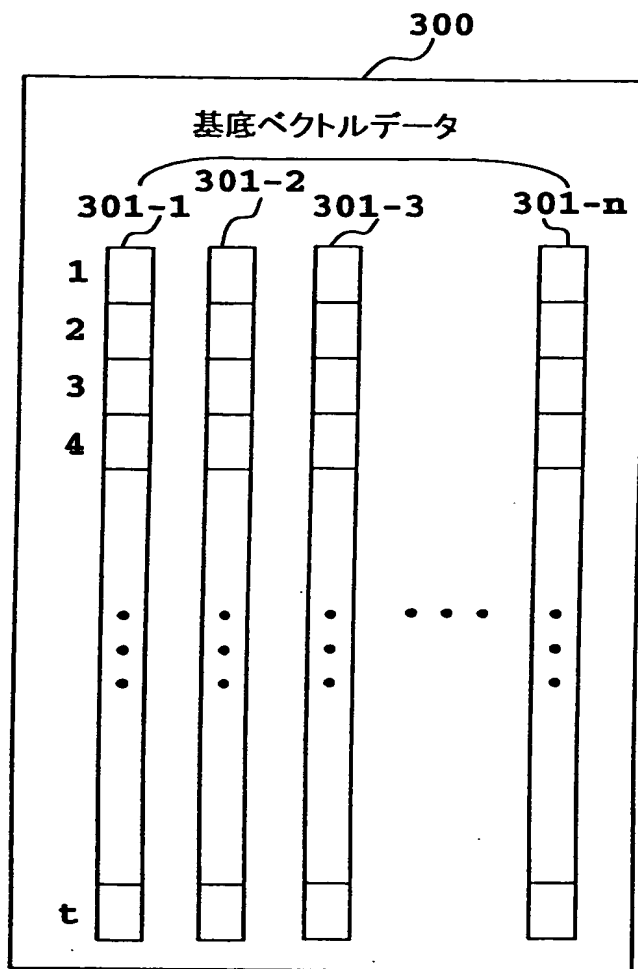
【図 6】



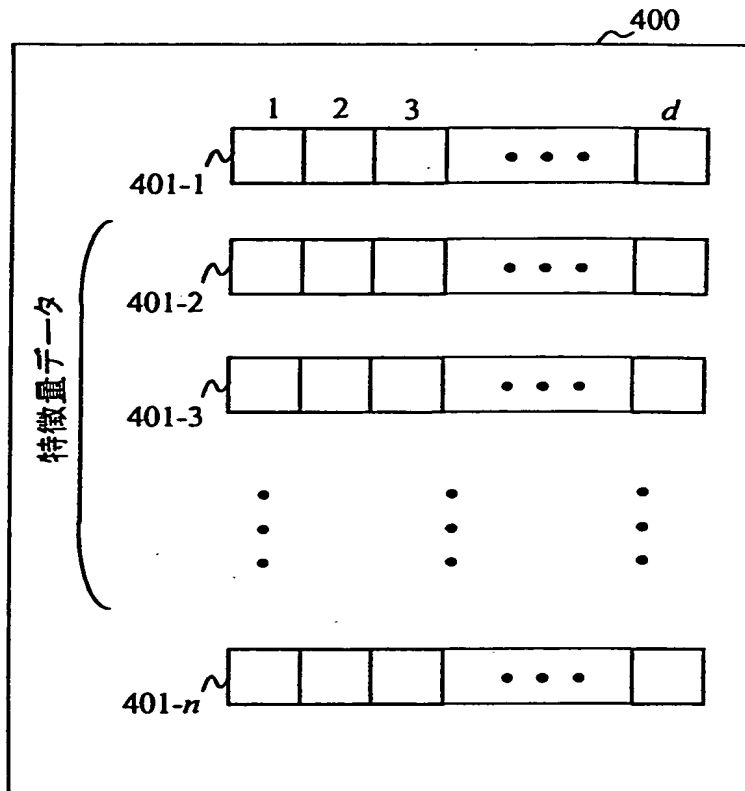
【図 7】



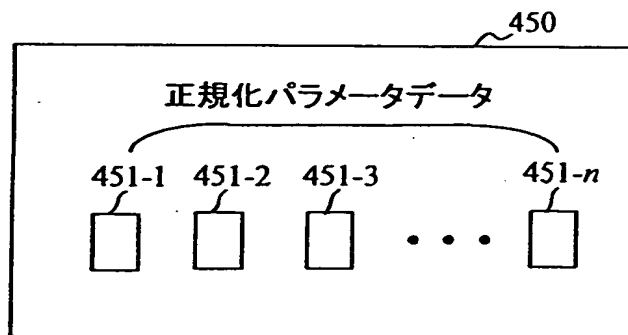
【図 8】



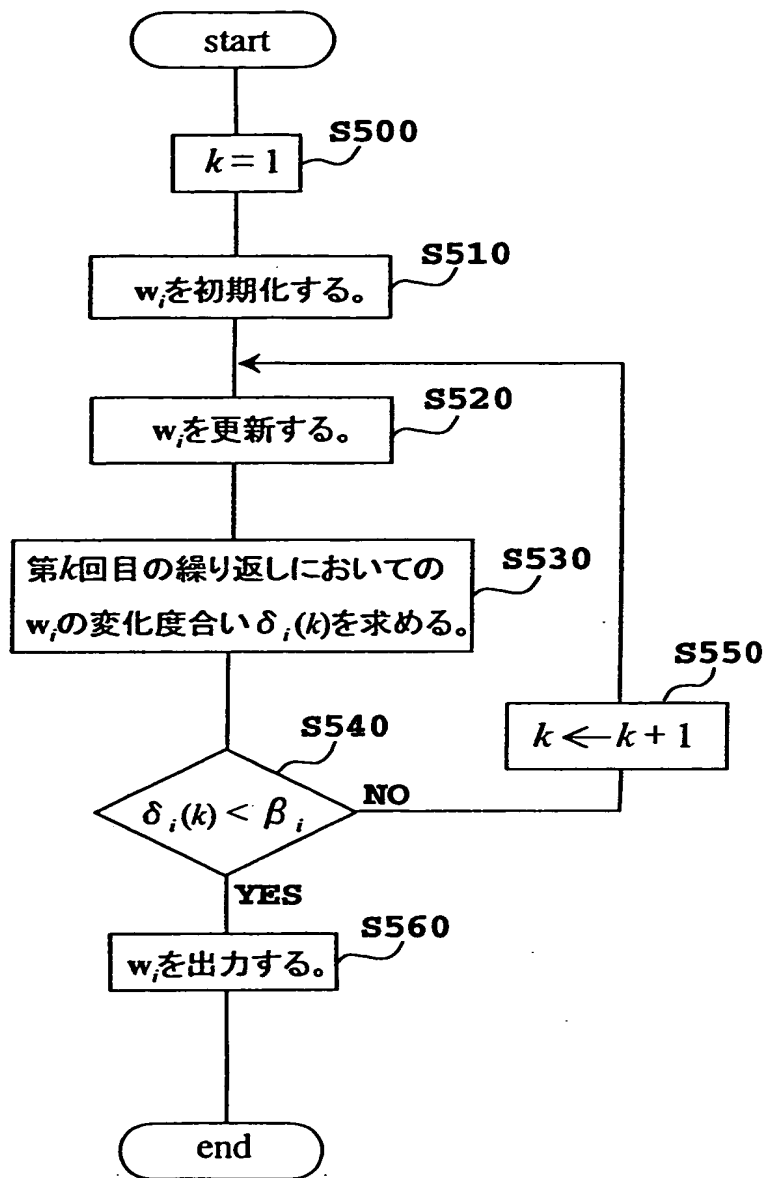
【図 9】



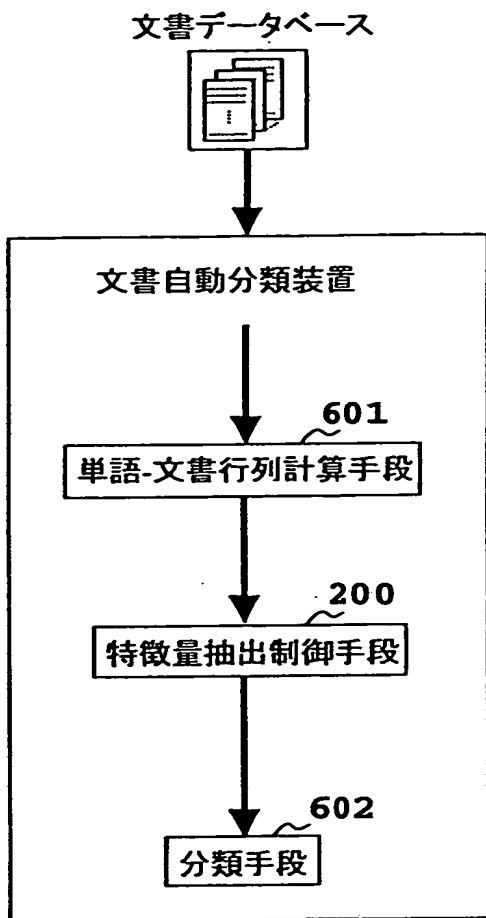
【図 1 0】



【図 1 1】



【図 1 2】



【書類名】 要約書

【要約】

【課題】 テキストマイニングにおける文書の特徴量抽出に関し、L S Aを実行可能な装置よりも小さいメモリ空間及びより簡単なプログラム及び装置で、L S Aと同じ性質を持つ特徴量を抽出する方法及びその装置を提供する。

【解決手段】 単語一文書行列更新手段 2 1 0 により更新された単語一文書行列、及びその行列を基に基底ベクトル計算手段 2 2 0 で計算された、有効な特徴量の空間を張る基底ベクトルを基に、特徴量抽出手段 2 3 0 で各文書の特徴量を抽出し、特徴量がユーザーにより与えられた所定の数を満たすまで各手段の実行を繰り返す。

【選択図】 図 5

出 願 人 履 歴 情 報

識別番号 [3 0 0 0 4 4 8 3 8]

1. 変更年月日 2 0 0 0 年 6 月 1 日

[変更理由] 新規登録

住 所 高知県南国市蛸が丘 1 - 1 - 1 南国オフィスパークセンター
3 F

氏 名 株式会社エス・エス・アール

出 願 人 履 歴 情 報

識別番号 [597154966]

1. 変更年月日 1997年11月 5日

[変更理由] 新規登録

住 所 高知県香美郡土佐山田町宮ノ口185番地
氏 名 学校法人高知工科大学